In: Fisher, R.B. and Hayes, G.M. (eds.) Proceedings of the 6th International Symposium on Intelligent Robotic Systems (SIRS'98), 21-23 July 1998, The University of Edinburgh, pp. 73-81.

Dynamic Expectancy: An Approach to Behaviour Shaping Using a New Method of Reinforcement Learning

Mark Witkowski[1]

Department of Computer Science,
Queen Mary and Westfield College (University of London),
Mile End Road,
London E1 4NS
United Kingdom

Abstract. This paper is concerned with issues relating to the source of reward and reinforcement with potential application to various robot learning and behaviour shaping situations (Dorigo and Colombetti, 1994; Lin, 1991, Maclin and Shavlik, 1996). The conventional approach to behaviour shaping by reinforcement learning is to present "reward" to an animal, animat or robot immediately following the performance by the animat of some required or desirable activity. It is a commonplace observation in experimental psychology that if this procedure is repeated a sufficient number of times by a trainer the behaviour of an animal will come to favour those activities in the circumstances under which they were reinforced.

This paper describes the *Dynamic Expectancy Model*, a new approach to issues in reinforcement learning that emphasises the role of internally generated "reward" signals, and in which overt behaviour is selected reactively from a policy map created dynamically in response to motivating "goals". The results of two investigations that illustrate these facets of learning and behaviour are presented. It is hoped that this technique will find application in a variety of task areas where animat/robot and man co-operate to address shared tasks.

1 Introduction

The recent introduction of systematic algorithms to propagate the effects of occasional reward has overcome a perennial problem with simple reinforcement, how to give the appearance of chained or goal-seeking behaviour. Such algorithms are exemplified by Watkins' *Q*-learning procedure (Watkins, 1989; Sutton, 1990; Watkins and Dayan, 1992), but there are many variants of the reinforcement procedure (Kaelbling *et al*., 1996, for review). The *Q*-learning algorithm effectively distributes the effects of external reward to generate a *static policy map*, from which the animat may select actions on the basis of current sensory information (*reactively*) to maximise the amount of reward obtained.

Cursory inspection of the *Q*-learning procedure reveals that no learning can occur until external reward is applied, and that learning and performance are intimately bound together, encoded in the policy map. The static policy map generated is highly specific to the source(s) of reward applied. Interest in practicable applications using reinforcement learning techniques as a robot learning and control architecture leads us to consider two related issues:

- Is external reward necessary for reinforcement learning?
- To what extent are reward and learning inter-related?

The middle part of the paper is given over to a description of the Dynamic Expectancy Model. The remainder of the paper then presents two investigations using the C++ implementation of the model (*SRS/E*) relevant to these questions. The first investigation replicates the established *latent learning* procedure (Bower and Hilgard, 1981; Riolo, 1991). Latent learning experiments were originally devised to refute the (then widely held) view that natural learning was entirely mediated by external reinforcement. Having shown this need not be the case, and replicated the essential conditions used in the original animal experiments with SRS/E, it is then necessary to explain why so much observable learning does indeed appear to be mediated by external reward. This is the subject of the second investigation.

[1] Email: markw@dcs.qmw.ac.uk

2 The Dynamic Expectancy Model

In contrast to the two-part (S-R or "stimulus-response") representations used by external reinforcement algorithms the Dynamic Expectancy Model is based on a three-part learned representation, the expectancy or *μ-hypothesis*. μ-Hypotheses are constructed from two *signs* ('s1' and 's2') and one *response* ('r1'). Signs are conjunctions of sensory *tokens*. A sign is defined as active when all the elements of the conjunction are satisfied. A response is selected from a number of predefined elementary motor actions or compound behaviours that the animat can perform. Recent active tokens, signs and responses are recorded in *traces*. Information in these traces is used to construct new μ-hypotheses as the need arises. Signs may also incorporate past tokens recorded in the trace as items in the conjunction. The form of the μ-hypothesis is:

$$\text{μ-hypothesis: } s1 + r1 \rightarrow^{t} s2 \qquad \text{(eqn. 1)}$$

A μ-hypothesis can be read as "performing the response 'r1' in the context defined by the sign 's1' predicts the occurrence of the sensory condition 's2' at time *t* in the future". A μ-hypothesis is itself said to be active, and so makes the prediction, whenever both the sign 's1' is active and the response 'r1' is being performed. Reinforcement is between the two items to the left of the '→', and the sign to its right.
Reinforcement, recorded in the *corroboration measure* (C_m) of each μ-hypothesis, is updated following every *activation* (prediction) of the μ-hypothesis. C_m is strengthened by the *reinforcement rate* α ($0 \leq \alpha \leq 1$) for each successful prediction, and weakened by the *extinction rate* β ($0 \leq \beta \leq 1$) according to:

$C_m = C_m + \alpha(1 - C_m)$ where a prediction is successful, or (eqn. 2a)
$C_m = C_m - \beta(C_m)$, where a prediction is unsuccessful, or (eqn. 2b)
Unchanged otherwise.

Reinforcement is thus fully independent of any external reward and may be made immediately the predictive expectancy has been validated. Predictions awaiting validation are stored within the model, but are discarded once the corroboration measure is updated.

2.1 Default Behaviours and Setting Goals

Normally the system selects responses reactively according to some fixed set of *behaviour rules*. Behaviour rules are pre-programmed and define appropriate responses for the robot either before learning has taken place, or in situations where learned responses are inappropriate (i.e. in highly safety critical situations where the variability introduced by learned behaviour is counter productive.) If no such behaviour rules are applicable or defined, responses are selected according to a default exploratory regime (the selection of responses at random being one option.)

Any sign known to the system can be designated a *goal*. Several goals may be asserted at any time, they are held on the *goal list* ordered by their *priority*. Only one, the *top-goal*, that of the highest priority, motivates the behaviour of the system at any one time. Once a top-goal is established the system switches to an exploitative or *valenced* mode of behaviour in which exploratory activities are suspended and purposive, goal-directed, behaviour is adopted. A goal is deemed satisfied when the sign that defines it becomes active. Satisfied goals are removed from the goal list, and remaining goals reordered according to their current priority. Goals may be asserted according to (a sub-set of) the pre-defined behaviour rules. Goals may also be asserted externally by an experimenter or operator. The investigations described in this paper use the external assertion method.

2.2 Constructing a Dynamic Policy Map

Whenever a top-goal is asserted the system attempts to construct a *Dynamic Policy Map* (DPM) from all the μ-hypotheses known to the system at the time by a process of *spreading activation*. The DPM takes the form of a graph in which the nodes are signs and the arcs transitions represented by individual μ-hypotheses. The top-goal node is said to have a *valence level* (*v*) of zero, any signs that may be reached in a single arc a valence level of one, and so on. The net effect of the DPM construction process is to characterise each sign (s1) at some level (*n*) according to a *policy value*, P_v:

$$P_v(s1_n) \leftarrow \min \left(\sum_{v=1}^{v=n} (C_e)_v \right) \qquad \text{(eqn. 3)}$$

This chain of nodes and arcs defines the sequence of signs and actions that are estimated to take the animat from the sign $s1_n$ to the goal with the minimum estimated cost. For each sign in the DPM there is one such *valenced path* (there may, of course, be many other paths.) C_e, the *cost estimate*, is calculated from the C_m value associated with the arc and the actual physical effort of performing the response associated with the arc:

$$C_e \leftarrow \text{response\_cost}(r1) / C_m \qquad \text{(eqn. 4)}$$

The *response cost* is a pre-defined measure of the effort required to perform the response (r1) associated with the μ-hypothesis. Response costs may be expressed in terms of energy expended, time taken or any other units that can be consistently applied to the responses defined in the system. Once constructed the DPM acts in a similar manner to a static policy map, allowing responses to be selected reactively (from the arc μ-hypothesis) solely on the basis of incoming sensory conditions. The animat selects responses from the valenced path of lowest estimated cost associated with one of the currently active signs.

The computational cost of constructing a complete DPM is broadly characterised by the relationship:

$$\tfrac{1}{2}H * V \qquad \text{(eqn. 5)}$$

Where H is the number of μ-hypotheses known to the system at the time and V the maximum valence level reached before the DPM is completed. Construction of the DPM is complete when no more μ-hypotheses can be incorporated into the graph and when no valenced paths of lower estimated cost through the graph are possible. The DPM may be utilised by the robot as soon as it incorporates at least one active sign (as a valenced response is then available) although policy paths of lower estimated cost may still be undiscovered. It may also be that the DPM, although complete, incorporates no currently active signs. In this case the system selects a response on the basis of the default behavioural strategies. In doing so it may be that an active sign on the DPM is then encountered and a valenced response may subsequently be selected.

An incomplete DPM may arise for a number of reasons. First, the robot may have had insufficient opportunity to learn the μ-hypotheses required to construct a solution. Second, there may be insufficient processing power available to complete the construction of the DPM before the next response must be selected. While each step in the construction is not computationally expensive, the spreading activation process is clearly sensitive to the number of μ-hypotheses held by the system. This may be managed in a number of ways. The system may acknowledge the processing bound, treating the unfinished DPM as it would an incomplete one. It may spread the computation over several "sense-act" cycles, at the risk of employing valenced responses that are out of date. It may alternatively define a strategy to restrict the number of μ-hypotheses held by the system at any time to match the resource available to utilise them. As the number of μ-hypotheses held also affects the effort required to manage outstanding predictions and corroborations the management of μ-hypotheses is considered in the next section. None of the investigations described in this paper are resource bound in this manner.

The spreading activation method employed here is a modified form of the well-established breadth-first graph search/construction algorithm (Nilsson, 1980). The DPM is recomputed when the top-goal changes, or when the available μ-hypotheses have changed materially. Valenced behaviour selected from the DPM is terminated once outstanding goal conditions are satisfied and behaviour selection reverts to the default strategy until a new goal is asserted.

2.3 Managing μ-Hypotheses

The system begins with no stored μ-hypotheses. New μ-hypotheses are created whenever a previously unknown sign is encountered (*creation by novelty* method), or by the unexpected appearance of a previously known, but unpredicted sign is encountered (*creation by unexpected event* method). A new μ-hypothesis is created from the novel (or unexpected) sign forming the 's2' component, and a sign 's1' and response 'r1' drawn respectively from the sign and response traces recording past events. The timing factor *t* (eqn. 1) is derived from timing information held in the traces. As learning by creation is a resource-bound activity, the probability that a μ-hypothesis will be created is regulated by the *learning probability rate*, λ ($0 \leq \lambda \leq 1$). This parameter determines whether any particular opportunity to create a new μ-hypothesis will be taken or passed over. Where $\lambda < 1$ the system will retain signs for which no corresponding μ-hypothesis exists. However at subsequent (unpredicted) appearances of the sign the opportunity again arises to create a μ-hypothesis to predict that sign by the unexpected event creation rule.

Following a period of creation and corroboration each sign may be predicted by μ-hypotheses that either (1) completely or nearly completely predict correctly, (2) only partially predict correctly, or (3) predict at or below the rate determined by chance. Clearly μ-hypotheses falling into the first group merit little further attention, and those in group three may be discarded. Those in the second group may indicate that the condition component of the μ-hypothesis ('s1') is underspecified, in effect giving rise to the perceptual aliasing problem (Chrisman, 1992). Such μ-hypotheses can be *differentiated* by adding further tokens to the 's1' component, taken either directly from the active tokens, or from the token trace if a temporal discrimination is required. This procedure is used to create competing μ-hypotheses, and introduces new signs into the system. To be effective these methods for μ-hypothesis management must take into account several factors, and several strategies have been considered (Witkowski, 1997).

3 Latent Learning

The latent learning procedure neatly demonstrates both the separation of learning from performance and distinguishes between the effects of "internal" and "external" reinforcement. The procedure compares the learned performance of an animal (or animat or robot in our case) in three distinct conditions:

1. Where the task is always rewarded
2. Where the task is initially unrewarded, but where reward is later introduced
3. Where the task is always unrewarded

In case 1 we expect the animat to both learn and hence improve task performance on subsequent trials, at some rate determined by the characteristics of the underlying learning mechanism. Typically this will reflect the negatively accelerating learning curve characteristic of this type of task (Sutton, 1990). After sufficient trials performance typically stabilises at or about the optimal performance level.

In case 2, if external reward is indeed required for learning, we would expect no learning to take place (as there is no external reward applied) and so no performance improvement would be apparent. If external reward is required to learn we would expect performance to begin to improve gradually following the availability of reward, much as it would do during case 1. Whereas if internal reward is being used learning takes place during the initial phase of case 2, but no performance improvement will be apparent, as the animat has no reason to perform the "task". Once reward is introduced performance will improve abruptly, a result of the learning that has taken place, but which was not previously made manifest.

In case 3, the control, we would expect no significant task improvement in either case, regardless of whether learning had or had not occurred, as the animat is unrewarded throughout the procedure.

3.1 The Tolman and Honzik Experiments

We replicate the essential conditions employed by Tolman and Honzik (Bower and Hilgard, 1981, pp. 338-340). In the original experimental procedure Tolman and Honzik tested three groups of food deprived rats in a maze apparatus. Group 1 ("Regularly rewarded", figure 1) were allowed to wander the maze once per day and obtained food reward on reaching the end location. Each animal was placed in the apparatus once per day before being return to their normal accommodation. Once the animal reached the end location it was prevented from re-entering the body of the maze by an arrangement of one-way doors. Group 2 ("No food reward until day 11") were allowed to traverse the maze, but on reaching the end location received no food reward. On the eleventh day group 2 were given food reward in the end location. Group 3, acting as control, ran the maze once per day with no food reward throughout the duration of the experiment.

Group 1, figure 1, displays a normal learning curve, performance improving gradually, much as would be expected from, say, a *Q*-learning based animat ("*Q*-animat"). Up to day 11, Group 2, without food motivation, perform similarly to the control group. With no known source of reward to drive learning, or to motivate behaviour, performance remains largely constant. This again would be expected from a *Q*-animat. The rapid performance improvement following the introduction of food to group 2 would certainly not be expected from a conventional external reinforcement learning based *Q*-animat. Note that the performance drops below that for Group 2, demonstrating both that learning is indeed not dependent on external reward, and that motivation (hunger) dramatically modifies behaviour when the opportunity arises (food at the end location).

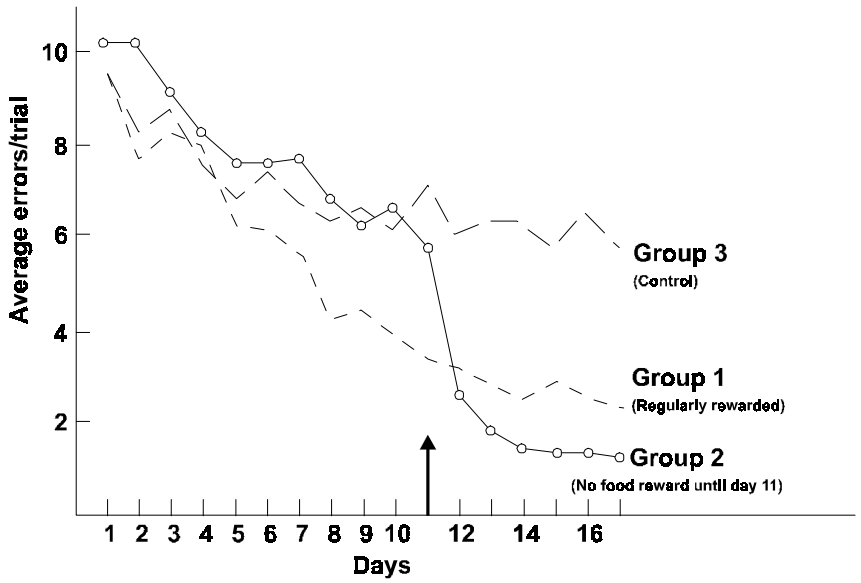

Figure 1: Tolman and Honzik's Latent Learning Results

3.2 Replicating the Latent Learning Experiments with SRS/E

Figure 3 shows the effects of running the latent learning experimental procedure described on an SRS/E based animat/robot ("SRS-animat"). In these experiments the robot is placed in the maze shown in figure 2 at the start point shown and allowed to traverse the maze until it reaches the end or "goal" point at the top of the maze, at which point the trial ends and the next may begin. The robot may recognise some 33 locations within the maze. These equate directly to signs. The robot is supplied with four responses with which to traverse the maze between signs. The SRS/E algorithm is directly interfaced to a Nomad robot (Nomadic Technologies, 1998) simulator, which provides a useful method to visualise the behaviour of the robot. The robot detects maze walls using the infrared range detectors ("short sensors", right of figure 2) provided. To conduct the bulk trials described in this section the robot may be disconnected and the effects of the maze traversal emulated. Where reward is required on any trial during the procedure the end location is asserted as a "goal" directly by the operator, i.e. that location is made "rewarding".

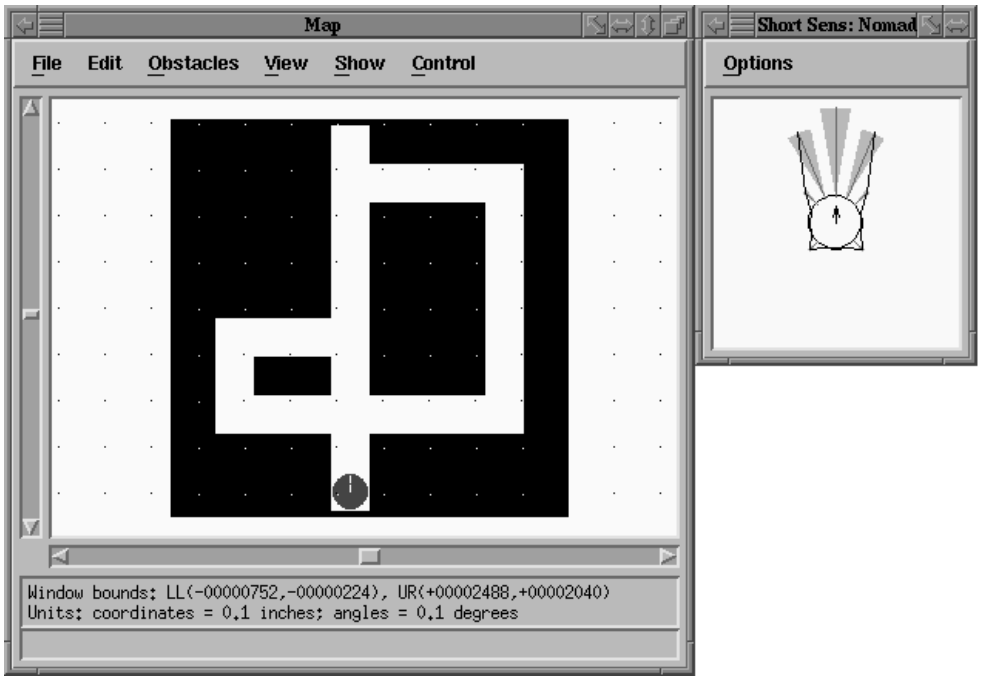

Figure 2: The "Nomad" Latent Learning Maze Layout

Each curve on figure 3 is the average performance through the maze apparatus of 100 individual animats. In this experiment $\alpha = 0.5$, $\beta = 0.2$ and $\lambda = 0.25$. Each group of animats is effectively "cloned", and so the first animats in each group behave identically until some condition changes, similarly the second and so on. Thus groups 2 and 3 track one another until trial 11, until the experimental conditions vary between the groups. Note the steady learning curve described by Group 1 ("Continuous reward"), and the abrupt change in behaviour for Group 2 ("No reward until trial 11") to near optimal performance (9 steps/trial) once reward is introduced.

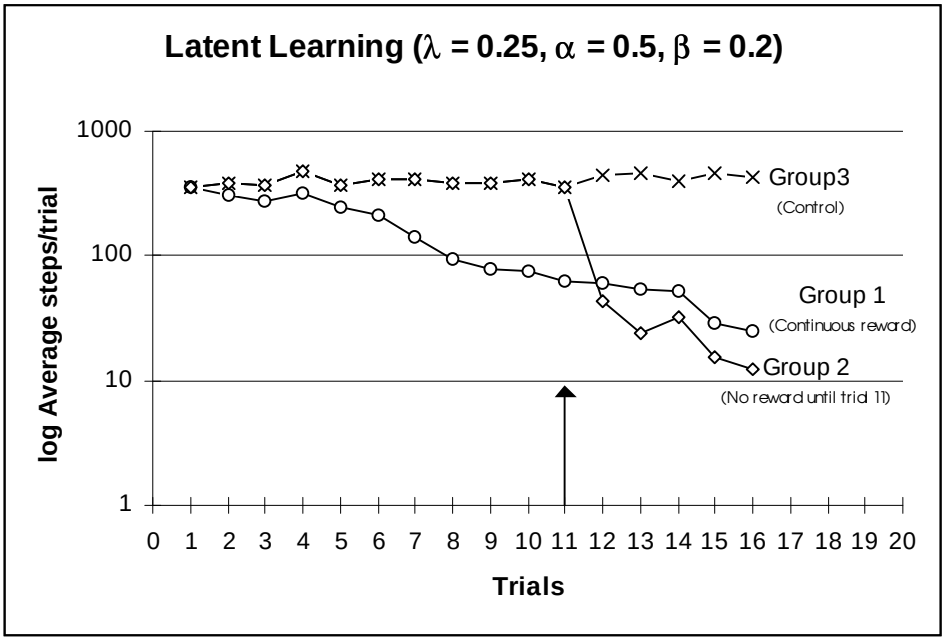

Figure 3: Results of the SRS/E Latent Learning Experiment

3.3 Latent Learning - Discussion

Latent learning phenomena have been widely investigated under a range of possible conditions (Thislethwaite, 1951). Not every investigation has successfully demonstrated the effect. However, several commentators have observed that any valid demonstration of latent learning must be deemed a strong challenge to the conventional view of reinforcement learning by external reward. Considerable care is needed to define and control the parameters to successfully and repeatably observe latent learning.

In the Dynamic Expectancy Model, as implemented by the program SRS/E, reinforcement learning by corroboration takes place at every step, regardless of the state of the animat's motivation or the availability of a "goal" condition to satisfy that motivational state. μ-Hypothesis formation takes place whenever the novelty or unexpected conditions arise, at a rate determined by the learning probability rate, λ.

During the first trial each animat in group 1 explores the maze using a random walk strategy. On each subsequent trial individual animats will acquire sufficient μ-hypotheses to build an effective DPM, at first partial, but eventually spanning between start to end location. In some cases this route between start and end will be via the shortest route. In the remaining cases the path will be one of the other two, longer, routes. Once a valenced path is established the animat will continue to use it in preference to exploring for new and possibly better (but possibly worse) routes while it remains continually valenced. Whilst we may only conjecture that the mechanism giving rise to these phenomena in the Dynamic Expectancy Model also applies to animals, such wilful "overtasking" is a pathological case for the Dynamic Expectancy Model, and may also be for animals and humans. In a normal course of events the motivation will be sated once the goal is achieved and the animat begin to pursue some other goal from the goal list or revert to an explicitly exploratory strategy. In doing so new μ-hypotheses will be formulated and existing ones corroborated to provide a different, and perhaps more effective, DPM when the original task is returned to at a later time.

The rate at which individuals achieve an effective DPM varies considerably, some doing so almost immediately, some taking many trials. The shape of the learning curve for group one, figure 3, is a direct consequence of averaging these different individual learning sequences. The slope of the Group 1 learning curve is primarily controlled by λ, becoming shallower as λ is reduced. Perhaps surprisingly, the reinforcement values α and β (eqns. 2) have little effect on the learning rate, but rather primarily determine which path the

animat will select through the DPM, when there is a choice. The SRS/E algorithm has been extensively tested with various values for λ (Witkowski, 1997) and the value used here (0.25) was selected to match the earlier findings. No doubt Tolman and Honzik's experimental design choices were based on their extensive knowledge of the natural learning rates exhibited by rats in these kinds of apparatus.

By freely using the random walk strategy to explore the maze until trial 12 group 2 animats build a more complete set of μ-hypotheses and corroborate them more thoroughly before being called upon to build the DPM and switch to valenced behaviour. As a consequence they demonstrate an apparently disproportional shift in performance when motivation is applied, because more individuals have located the optimal path (97% for group 2 as opposed to 87% for group 1 at this point) before valenced behaviour from the DPM is invoked.

The conditions used in the animal and those used in the animat experiments are not identical, each is adapted to the properties of the subject under test. However, the essential demonstration of the latent learning phenomena is clearly seen in both, the use of the log scale on the "steps/trial" axis of figure 3 is largely cosmetic and compensates for the longer path traversals encountered in the emulation. Tolman and Honzik reportedly employed a maze constructed from 14 'T'-shaped choice points. In each case the arm of the 'T' being blocked to form a 'cul', the other choice continuing toward the end location. One way doors were fitted between segments in this maze to prevent the animal returning along its previous path. Performance in the maze could be measured as the number of "errors" (turns into the blocked arms) made in traversing from start to finish. In the emulation results presented here the three-path maze of figure 2 is adopted, with no additional restrictions to the direction the robot may make within the confines of the maze.

With the clarity of hindsight it is easy to see that it is the existence of shorter and longer paths which become fixed as a result of the goal-seeking, coupled to careful management of reward and motivation, explains the particular cross-over effect between the performance in groups 1 and 2. The latent learning phenomena has been tested using the Dynamic Expectancy Model with several designs of experimental apparatus, and has been found to be reliably repeatable.

Closer inspection of figure 1 reveals two additional phenomena worthy of comment. First, both group 2 ("no reward until day 11") and group 3 ("control") show some performance improvement even when no reward is presented. It may be that the animals found the act of being removed from the end location and returned to their cages in some way "rewarding", perhaps through some prior association with the availability of food in their normal accommodation. Tolman and Honzik left the subject animals in the end location for some time at the conclusion of each trial in an attempt to overcome this potential experimental difficulty. It might also be that rats have a natural disposition to ignore short, blocked paths (which, given their ancestral habitats, would not be unreasonable); or be related to a more generalised "curiosity" penchant, which is frequently observed in mammals and biases them to explore novel situations. No such mechanisms are incorporated in the Dynamic Expectancy Model and the effect is not apparent in figure 3.

Second, group 2 (figure 1, no reward until day 11) demonstrates a consistent level of errors even after the reward regime is established. It may, of course, be that rats have poor memory or recall; but it may arise as a consequence of a broader inherent behavioural strategy to check previously learned knowledge to specifically check for better solutions. This interpretation is a manifestation of the *exploration-exploitation tradeoff*, which is an integral component of conventional reinforcement learning algorithms. With the Dynamic Expectancy Model "exploration" and "exploitation" are kept largely separate due to the explicit alternation between innate and valenced activities, although learning, both by the creation of new μ-hypotheses and the corroboration of existing ones, may occur at any time. Preliminary investigations in which an element of exploration is reintroduced in valenced behaviour indicate that the disruption to the existing valenced solution outweighs the advantage gained by the extra opportunity to learn better strategies. It may be that further investigation will reveal an overall effective balance. No exploration bias used in the results shown in figure 3 during valenced behaviour.

4 Combining Reinforcement with Motivation

Latent learning experiments demonstrate that both in animals and, with the Dynamic Expectancy Model, in animats or robots there can be reinforcement learning without external reward. However, it is clear that external reward is an important factor in shaping animat behaviour. It would be idle to suggest the animat should always perform learning for its own sake, learning must for the most part be biased towards the needs and motivations of the animat, as has been consistently demonstrated to be the case in natural learning studies.

This bias towards learning behaviours directly related to the animat's motivators (its goals) is achieved in the Dynamic Expectancy Model in a process referred to as *valence level pre-bias*. The model separates the quality of an observation from its usefulness, where conventional reinforcement learning algorithms do not.

The corroboration measure, C_m, for each μ-hypothesis is calculated as before, rightly reflecting its viability as a predictive expectancy. Instead the probability that a rule is created (λ) is varied, predisposing the algorithm to create potentially useful μ-hypotheses. A new measure, the *effective learning probability rate* (λ'), is derived from λ and the best (lowest) valence level recorded for each sign. λ' is subsequently used in preference to λ to determine whether a new μ-hypothesis will be formed, when opportunities arise to use either of the μ-hypothesis creation methods previously described. μ-Hypotheses creation is therefore biased according to the extent to which their component signs were implicated in previously valenced paths. Figure 4 presents experimental results that show the significant performance gains in learning that may be demonstrated when internal reinforcement and motivation are combined in this manner. Biasing μ-hypothesis formation in this manner concentrates learning resource towards those signs which have previously been implicated in a Dynamic Policy Map and allows an effective DPM to be formed more quickly.

An alternative approach to the external reinforcement issue is to increase the effective value of the reinforcement measure, α, whenever external reward is applied, magnifying the effect of internal corroborative reinforcement. The effect of this tactic is to favour use of these μ-hypotheses for selection from the DPM as the effective cost estimate, C_e (eqn. 4), is lowered. As a side-effect these "rewarded" μ-hypotheses will be favoured in all subsequent constructions of the DPM, whether or not that μ-hypotheses is indeed relevant to the top-goal currently being pursued. It may well be that animals do indeed combine these effects, it is less clear whether this is desirable in the case of an animat or robot controller.

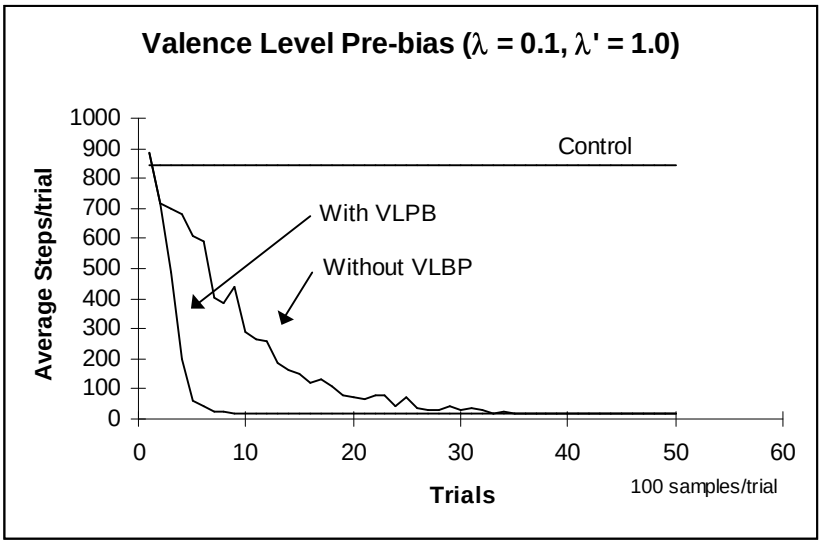

Figure 4: The Effect of Valence Level Pre-bias (VLPB)

5 Robot Behaviour Shaping – an Application Domain for the Dynamic Expectancy Model

Although much of the design of the Dynamic Expectancy Model is derived from experiments in animal learning and behaviour, and its implementation by the techniques of machine learning, it is expected that it will find application in the area of co-operative working between human and robot, the *behaviour shaping* domain. In this domain the human operator guides or commands the robot remotely, using joysticks or other forms of input, determining the responses the robot will use. This overrides the pre-programmed or default behaviours previously described. The robot still receives the stream of input tokens and signs from its sensors and will create and corroborate μ-hypotheses in the normal way. The learning mechanism is insensitive to the source of signs and responses.

At any point the human operator may select any sign as a goal, causing a DPM to be created. If the operator is satisfied that this represents an appropriate course of action he or she relinquishes control to the model until the task is completed. At any point the operator may interrupt the valenced behaviour of the robot to shape future behaviours without recourse to notions of overt external reward or punishment. In this way the experience of the robot, from which it learns, is channelled by the actions of the operator avoiding the need for explicit programming and overcoming the (probably wasteful) effort inherent in random exploration. In defining goals the operator may further focus robot learning to useful areas.

6 Concluding Remarks

This paper has described the *Dynamic Expectancy Model*, a design for an animat or robot controller that adopts and adapts two core ideas from the field of reinforcement learning. First that "reward" enhances the likelihood that actions or behaviours are selected reactively from a *policy map*, which is constructed iteratively by propagating the effects of external reward. It differs radically from the prevailing model of reinforcement learning in two significant ways:

1. "Reward" is primarily generated internally by the corroboration of predictive expectancies ("μ-hypotheses"), rather than as an imposed or external input.
2. A policy map (the *Dynamic Policy Map*) is created as required from the μ-hypotheses only when a specific task motivates the system (the *top-goal*). When the top-goal is satisfied the DPM is discarded, if the top-goal changes a new DPM specific to that new goal is created. The change in overt behaviour shifts immediately from one policy to the next.

The latent learning procedure described in the paper serves to highlight the differences between internal reinforcement by corroboration of expectancies, and external reinforcement by reward. The Dynamic Expectancy Model also derives much from models of "intermediate level cognition" (Becker, 1973; Mott, 1981; Drescher, 1991) and attempts a synthesis of the best features from the cognitive and reinforcement/reactive points of view. The Dynamic Expectancy Model and SRS/E are described in greater detail in Witkowski (1997), which also reports an extensive series of experiments and trials with the system, including the latent learning procedure described here.

References

Becker, J.D. (1973) "An Information-processing Model of Intermediate-level Cognition", in: Schank, R.C. and Colby, K.M. (Eds.) "Computer Models of Thought and Language" W.H. Freeman & Co., pp. 396-434

Bower, G.H. and Hilgard, E.R. (1981) "Theories of Learning", Prentice Hall (5th edition)

Chrisman, L. (1992) "Reinforcement Learning with Perceptual Aliasing: The Perceptual Distinctions Approach", in: Proc. of the American Association for Artificial Intelligence (AAAI-92), pp. 183-188

Drescher (1991) "Made-up Minds: A Constructivist Approach to Artificial Intelligence", MIT Press

Dorigo, M. and Colombetti, M. (1994) "Robot Shaping: Developing Autonomous Agents Through Learning", Artificial Intelligence, **71**, pp. 321-370

Lin, L-J (1991) "Programming Robots Using Reinforcement Learning and Teaching", in: Proc. of the American Association for Artificial Intelligence (AAAI-91), pp. 781-786

Kaelbling, L.P., Littman, M.L. and Moore, A.W. (1996) "Reinforcement Learning: A Survey", Journal of Artificial Intelligence Research, **4**, pp. 237-285

Maclin, R. and Shavlik, J.W. (1996) "Creating Advice-taking Reinforcement Learners", Machine Learning, **22**, pp. 251-282

Mott, D.H. (1983) "Sensori-motor Learning in a Mobile Robot", Dept. Comp. Sci., Queen Mary College, University of London, Ph.D. Thesis

Nilsson, N.J. (1980) "Principles of Artificial Intelligence", Springer-Verlag (Symbolic Computation Series)

Nomadic Technologies (1998) "Nomadic Technologies: Merging Mind and Motion", http://www.robots.com

Riolo, R.L. (1991) "Lookahead Planning and Latent Learning in a Classifier System", Proc. 1st Int. Conf. on Simulation of Adaptive Behavior "SAB-1", pp. 316-326

Sutton, R.S. (1990) "Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming", in: "Proc. 7th Int. Conf. on Machine Learning", Morgan Kaufmann, pp. 216-224

Thistlethwaite, D. (1951) "A Critical Review of Latent Learning and Related Experiments", Psychological Bulletin, **48**-2, pp. 97-129

Watkins, C.J.C.H. (1989) "Learning from Delayed Rewards", King's College, Cambridge Univ. (Ph.D. thesis)

Watkins, C.J.C.H. and Dayan, P. (1992) "Technical Note: *Q*-learning", Machine Learning, **8**, pp.279-292

Witkowski, C.M. (1997) "Schemes for Learning and Behaviour: A New Expectancy Model", Department of Computer Science, Queen Mary & Westfield College (University of London), Ph.D. thesis, February 1997